MAS187/AEF258

University of Newcastle upon Tyne

2005-6

Contents

Chapter 1

Collecting and Presenting Data

1.1 Introduction

Data are the key to many important management decisions. Is a new product selling well? Do potential customers like the new advertising campaign? These are all questions that can be answered with data. We begin this course with some basic methods of collecting, representing and describing data. In this first lecture we will look at the different kinds of data that exist, how we might obtain the data and basic methods for presenting them.

1.1.1 Examples

Sizing Clothes Most clothing now comes in essentially standard sizes but from where do these standards come? By sampling from the general population as a whole, standards can be set around the most common sizes. We can not say that an individual is exactly a standard size. However we can say that they will probably fall within a range either side of a standard.

Car Maintenance If we were buying a new car, it would be useful to know how much it was going to cost to run it over the next three years. Obviously we can not predict this exactly as each individual car and each user will be slightly different. Collecting data from people who have bought similar cars will give us some idea of the distribution of costs over the population of car buyers, which in turn provides us with information as to the likely cost of running the car.

1.1.2 Definitions

The quantities measured in a study are called *random variables* and a particular outcome is called an *observation*. A collection of observations is the *data*. The collection of all possible outcomes is the population.

If we were interested in the height of people doing management courses at Newcastle, that would be our random variable, a particular person's height would be the observation and if we measured

everyone doing MAS187, those would be our data, which form a sample from the population of all students registered with the School of Management.

In practice it is difficult to observe whole populations, unless we are interested in a very limited population, e.g. the students taking MAS187. In reality we are usually observe a subset of the population, we will come back to sampling later in section 1.2.

Variables are of two types, *qualitative* and *quantitative*. Qualitative variables have non-numeric outcomes. They are usually *categorical*. Examples of qualitative variable include sex of a person or animal, colour of a car, mode of transport, football team supported. Quantitative variables have numeric outcomes with a natural ordering. Examples include people's height, time to failure of a component, number of defective components in a batch.

Quantitative variables are usually of one of two types: *discrete* or *continuous*. Discrete random variables can only take a sequence of distinct values which are usually the integers, although not necessarily so. Discrete variables are countable, for example the number of defective pieces in a manufacturing batch, the number of people in a tutorial group, or a person's shoe size. There are other kinds of discrete data. *Ordinal* data are data are ordered but which are not really numbers in the usual sense. For example, if you are asked to rank a response to a question between 1 and 10, from strongly disagree to strongly agree, an answer of 8 obviously indicates stronger agreement than one of 4, but not necessarily twice as strong in any meaningful sense.

Continuous variables can take any value over some continuous scale. Simple examples include height, weight, time taken to be served in a bank queue or the fuel consumption of a car. The important thing to note about continuous data is that, no matter how small an interval we consider, it is always possible (in theory, at least) to make an observation in the interval by using sufficiently precise measurement. We can measure to differing degrees of accuracy using different equipment but we could never say absolutely precisely how much someone weighs. Continuous variables are often expressed up to a number of significant digits and could appear to be discrete. It is the underlying variable which defines their status and not the form in which they are expressed.

1.1.3 Surveys

Surveys or questionnaires are often used to gain insight into the impact of many management decisions. For example, the market prospects of a new product or customer views on the impact and potential of the new technologies. When preparing a survey there are many key questions to consider:

- what is the purpose of the survey?
- what is the target population?
- is there a list of the target population?
- how can bias be avoided?
- how accurate does the survey have to be?
- what resources are available for conducting the survey?

- how are the data to be collected?

There are many ways of collecting survey information. Each has its advantages and disadvantages regarding the cost of implementation, the response rate (of successfully completed questionnaires), the speed with which the survey can be completed and the quality and accuracy of the information collected. The three main ways are:

Postal questionnaire – low cost, low response rate, slow turn around time, low quality information

Telephone interviewing – moderate cost, moderate response rate, fast turn around time, good quality information (?)

Face-to-face interviewing – high cost, high response rate, fast turn around time, high quality information

1.2 Sampling

We can rarely observe the whole population. Instead we observe some sub-set of this called the sample. The difficulty is in obtaining a *representative* sample. For example if you were to ask the people leaving a gym if they took exercise this would produce a *biased* sample and would not be representative of the population as a whole. The importance of obtaining a representative sample can not be stressed too highly. As we will see later we use the data from our samples in order to make inferences about the population and these inferences influence the decision making process.

There are three general forms of sampling techniques.

Random sampling where the members of the sample are chosen by some random mechanism.

Quasi-random sampling where the mechanism for choosing the sample is only partly random.

Non-random sampling where the sample is specifically selected rather than randomly selected.

1.2.1 Simple Random Sampling

This method is the simplest to understand. If we had a population of 200 students we could put all their names into a hat and draw out 20 names as our sample. Each name has an equally likely chance of being drawn and so the sample is completely random. Furthermore each possible sample of 20 has an equal chance of being selected. In reality the drawing of the names would be done by a computer and the population and samples would be considerably larger.

The disadvantages of this method are that we often do not have a complete list of the population. For example if you were surveying the market for some new software, the population would be everybody with a compatible computer. It would be almost impossible to find this information out. Not all elements of the population are equally accessible and hence you could waste time and money trying to obtain data from people who are unwilling to provide it. Thirdly it is possible that purely by chance you could pick an unrepresentative sample, either over or under representing elements of the population. Using our software example you could pick by chance only companies that have recently updated their software and hence would not be interested in your new package.

1.2.2 Stratified Sampling

This is a form of random sample where clearly defined groups, or *strata*, exist within the population, for example males and females, working or not working, age groups etc. If we know the overall proportion of the population that falls into each of these groups, we can randomly sample from each of the groups and then adjust the results according to the known proportions. For example, if we assume that the population is 55% female and 45% male and we wanted a sample of 1000. We would first decide to have 550 females and 450 males in our sample. We would then pick the members of our sample from their respective groups randomly. We do not have to make the numbers in the samples proportional to the numbers in the strata because we can adjust the results but sampling within each stratum ensures that that stratum is properly represented in our results and gives us more precise information about the population as a whole. Such sampling should generally reflect the major groupings within the population.

The disadvantages are that we need clear information on the size and composition of each group or stratum which can be difficult to obtain. We still need to know the entire population so as to sample from it.

1.2.3 Systematic Sampling

This is a form of quasi-random sampling which can be used where the population is clearly structured. For example you were interested in obtaining a 10% sample from a batch of components being manufactured, you would select the first component at random, after that you pick every tenth item to come off the production line. This simplicity of selection makes this a particularly easy sampling scheme to implement, especially in a production setting.

The disadvantages of this method are that it is not random and if there is a pattern in the process it may be possible to obtain a biased sample. It is only really applicable to structured populations.

1.2.4 Multi-stage Sampling

This is another form of quasi-random sampling. These types of sampling schemes are common where the population is spread over a wide geographic areas which might be difficult or expensive to sample from. Multi-stage sampling works, for example, by dividing the area into geographically distinct area, randomly selecting one of these areas and then sampling, whether by random, stratified or systematic sampling schemes within this area. For example, if we were interested in sampling school children, we might take a random (or stratified) sample of education authorities, then, within each selected authority, a random (or stratified) sample of schools, then, within each selected school, a random (or stratified) sample of pupils. This is likely to save time and cost less than sampling from the whole population.

The sample can be biased if the stages are not carefully selected. Indeed the whole scheme needs to be carefully thought through and designed to be truly representative.

1.2.5 Cluster Sampling

This is a method of non-random sampling. For example, a geographic area is sub-divided into clusters and all the members of the cluster are then surveyed. This differs from multi-stage sampling covered in section 1.2.4 where the members of the cluster were sampled randomly. Here no random sampling occurs. The advantage of this method is that, because the sampling takes place in a concentrated area, it is relatively inexpensive to perform.

The very fact that small clusters are picked to allow the entire cluster to be surveyed introduces the strong possibility of "bias" within the sample. If you were interested in the take up of organic foods and were sampling via the cluster method you could easily get biased results, if for example you picked an economically deprived area, the proportion of those surveyed that ate organically might be very low, while if you picked a middle class suburb the proportion is likely to be higher than the overall population. (Technically, this is not strictly *bias* but *innefficiency* but, for now, it should be clear that there is a problem).

1.2.6 Judgemental sampling

This is an entirely non-random method of sampling. The person interested in obtaining the data decides whom they are going to ask. This can provide a coherent and focused sample by choosing people with experience and relevant knowledge to provide their opinions. For example the head of a service department might suggest particular clients to survey based on his judgement. They might be people he believes will be honest or have strong opinions.

This methodology is non-random and relies on the judgement of the person making the choice and hence it can not be guaranteed to be representative. It is prone to bias.

1.2.7 Accessibility sampling

Here the most easily accessible individuals are sampled. This is clearly prone to bias and only has convenience and cheapness in its favour. For example, a sample of grain taken from the top of a silo might be quite unrepresentative of the silo as a whole in terms of moisture content.

1.2.8 Quota Sampling

This method is similar to stratified sampling but uses judgemental (or some other) sampling rather than random sampling within groups. We would classify the population by any set of criteria we choose to sample individuals and stop when we have reached our quota. For example if we were interested in the purchasing habits of 18-23 year old male students, we would stop likely candidates in the street, if they matched the requirements we would ask our questions until we had reached our quota of 50 such students. This type of sampling can lead to very accurate results as it is specifically targeted, which saves time and expense.

The accurate identification of the appropriate quotas can be problematic. This method is highly reliant on the individual interviewer selecting people to fill the quota. If this is done poorly bias

can be introduced into the sample.

1.2.9 Sample Size

When considering collecting data, it is important to ensure that the sample contains a sufficient number of members of the population for adequate analysis to take place. Larger samples will generally give more precise information about the population. Unfortunately, in reality, questions of expense and time tend to limit the size of the sample it is possible to take. For example, national opinion polls often rely on samples in the region of 1000.

1.3 Frequency Tables

Once we have collected our data, often the first stage of any analysis is to present them in a simple and easily understood way. Tables are perhaps the simplest means of presenting data. There are many types of tables. For example, we have all seen tables listing sales of cars by type, or exchange rates, or the financial performance of companies. These types of tables can be very informative. However they can also be difficult to interpret, especially those which contain vast amounts of data.

Frequency tables are amongst the most common tables used and perhaps the most easily understood. They can be used with continuous, discrete, categorical and ordinal data. Frequency tables have uses in some of the techniques we will see in the next lecture.

1.3.1 Frequency Tables

The following table presents the modes of transport used daily by 30 students to get to and from University.

Student	Mode	Student	Mode	Student	Mode
1	Car	11	Walk	21	Walk
2	Walk	12	Walk	22	Metro
3	Car	13	Metro	23	Car
4	Walk	14	Bus	24	Car
5	Bus	15	Train	25	Car
6	Metro	16	Bike	26	Bus
7	Car	17	Bus	27	Car
8	Bike	18	Bike	28	Walk
9	Walk	19	Bike	29	Car
10	Car	20	Metro	30	Car

The table obviously contains much information. However it is difficult to see which method of transport is the most widely used. One obvious next step would be to count the number of students using each mode of transport:

Mode	Frequency
Car	10
Walk	7
Bike	4
Bus	4
Metro	4
Train	1
Total	30

This gives us a much clearer picture of the methods of transport used.

Also of interest might be the *relative* frequency of each of the modes of transport. The relative frequency is simply the frequency expressed as a proportion of the total number of students surveyed. If this is given as a percentage, as here, this is known as the *percentage relative frequency*.

Mode	Frequency	Relative Frequency (%)
Car	10	33.3
Walk	7	23.4
Bike	4	13.3
Bus	4	13.3
Metro	4	13.3
Train	1	3.4
Total	30	100

The data presented in the tables above are, of course, categorical. However other forms of data can also be presented in frequency tables. The following table shows the raw data for car sales at a new car showroom over a two week period in July.

Date	Cars Sold	Date	Cars Sold
01/07/04	9	08/07/04	10
02/07/04	8	09/07/04	5
03/07/04	6	10/07/04	8
04/07/04	7	11/07/04	4
05/07/04	7	12/07/04	6
06/07/04	10	13/07/04	8
07/07/04	11	14/07/04	9

Presenting these data in a relative frequency table by number of days on which numbers of cars were sold, we get the following table:

Cars Sold	Tally	Frequency	Relative Frequency %
Totals			

1.3.2 Continuous Data Frequency Tables

With discrete data and especially with small data sets it is easy to count the quantities in the defined categories. With continuous data this is not possible. Strictly speaking, no two observations are precisely the same. With such observations we group the data together. For example the following data set represents the service time in seconds for callers to a credit card call centre.

214.8412	220.6484	216.7294	195.1217	211.4795
195.8980	201.1724	185.8529	183.4600	178.8625
196.3321	199.7596	206.7053	203.8093	203.1321
200.8080	201.3215	205.6930	181.6718	201.7461
180.2062	193.3125	188.2127	199.9597	204.7813
198.3838	193.1742	204.0352	197.2206	193.5201
205.5048	217.5945	208.8684	197.7658	212.3491
209.9000	197.6215	204.9101	203.1654	192.9706
208.9901	202.0090	195.0241	192.7098	219.8277
208.8920	200.7965	191.9784	188.8587	206.8912

To produce a continuous data frequency table we first need to divide the range of the variable into smaller ranges called *class intervals*. The class intervals should, between them, cover every possible value. There should be no gaps between the intervals. One way to ensure this is to include the boundary value as the smallest value in the next class above. This can be written as for example, $20 \leq \text{obs} < 30$. This means we include all observations (represented by "obs") within this class interval that have a value of at least 20 up to values just below 30. Often for simplicity we would write the class intervals up to the number of decimal places in the data and avoid using the inequalities. For example 20 up to 29.999 if we were working to 3 decimal places. We need also to include the full range of data in our table and so we need to identify the minimum and

maximum points. (Sometimes our last class might be “greater than such and such”). Thirdly the class interval width should be a convenient number, for example 5, 10, 100 depending on the data. Obviously we do not want so many classes that each one has only one or two observations in it. The appropriate number of classes will vary from data set to data set. However, with simple examples that you would work through by hand, it is unlikely that you would have more than ten to fifteen classes. Bearing this in mind, let us create a frequency table for these data. As with discrete data frequency tables, we might also be interested in the percentage relative frequency of each class. This is simply calculated by taking the number in the class, dividing it by the total number in the sample and then multiplying this by 100% to obtain a percentage.

The data above give the following frequency table.

Class Interval	**Tally**	**Frequency**	**Relative Frequency %**
Totals			

1.4 Exercises 1

Identify the type of data described in each of the following examples.

1. An opinion poll was taken asking people which party they would vote for in a general election.
2. In a steel production process the temperature of the molten steel is measured and recorded every 60 seconds.
3. A market researcher stops you in Northumberland Street and asks you to rate between 1 (disagree strongly) and 5 (agree strongly) your response to opinions presented to you.
4. The hourly number of units produced by a beer bottling plant is recorded.

The following table includes data for the number of telephone call made by 50 students in a month.

98	99	99	100	100
101	100	104	97	101
102	100	99	101	99
100	96	99	101	99
99	98	95	99	99
97	101	100	101	101
103	102	96	98	103
98	100	102	99	101
98	99	100	98	99
102	98	99	99	97

Put these data into a relative frequency table.

The following data are the recorded length (in seconds) of 50 mobile phone calls made by one student. Construct a frequency table appropriate for these data.

281.4837	293.4027	306.5106	286.6464	298.4445
312.7291	327.7353	311.5926	314.8501	303.3484
270.7399	293.9364	310.9137	346.4497	304.6044
304.1124	320.7182	283.6594	337.5806	259.6408
305.4378	317.9180	289.5667	286.9626	300.5140
278.3108	300.1725	292.6725	312.9645	302.5770
293.2735	267.5344	326.9056	257.7226	285.9805
299.6535	293.9145	303.9191	323.7993	263.5242
281.1613	306.9344	310.2583	301.6963	313.9611
314.8500	292.0031	302.4314	267.9781	292.0917